

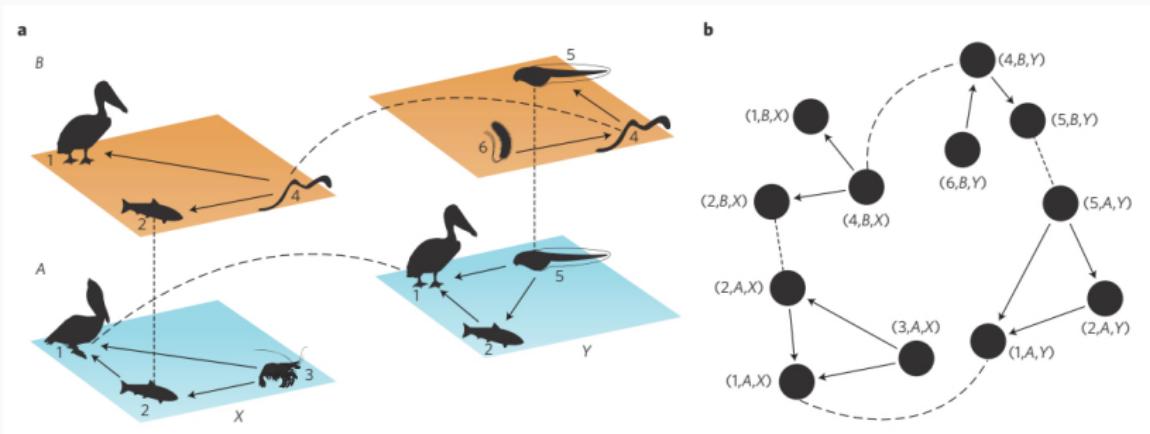
Tracking the data provenance for multilayer network modeling and analysis of foodwebs

Matthew K. Lau, Thomas F.J.M. Pasquier, Aaron M. Ellison (Harvard University)

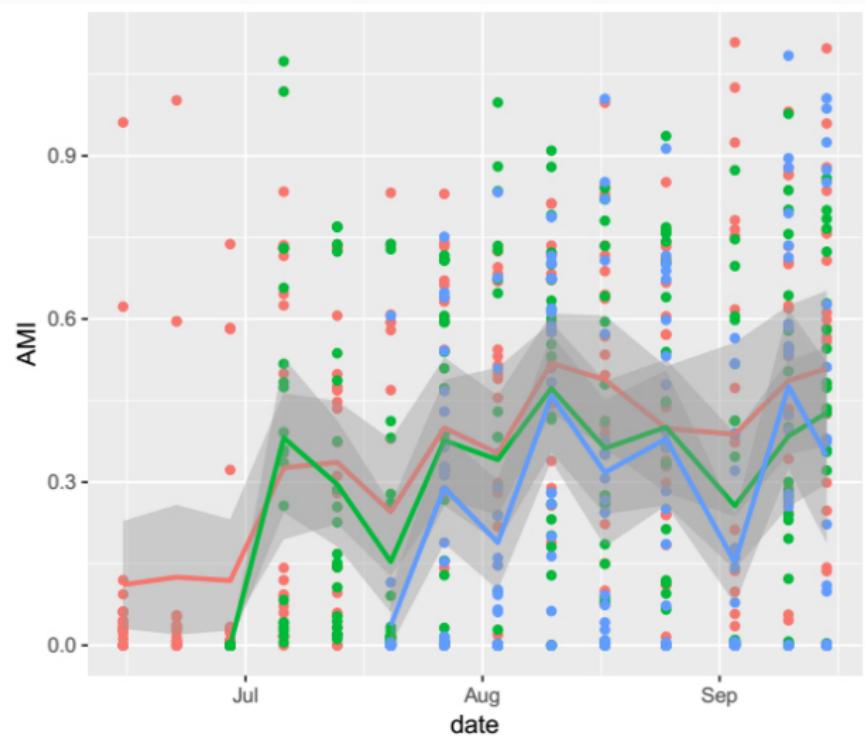
Pitcher Microecosystems



Multilayer network analysis



Multilayer networks



Data availability. The raw data for the example temporal network are in Supplementary Data 1 and deposited in Figshare (<https://dx.doi.org/10.6084/m9.figshare.3472646.v2>). The code for general procedures to prepare data, manipulate networks, post-process modularity-maximization calculations, and analyse network robustness are written in R, and they are available at <https://dx.doi.org/10.6084/m9.figshare.3472664.v1>. The code for examination of modular structure is written in MATLAB and available at <https://dx.doi.org/10.6084/m9.figshare.3472679.v1>.

Received 20 April 2016; accepted 27 January 2017;
published 23 March 2017

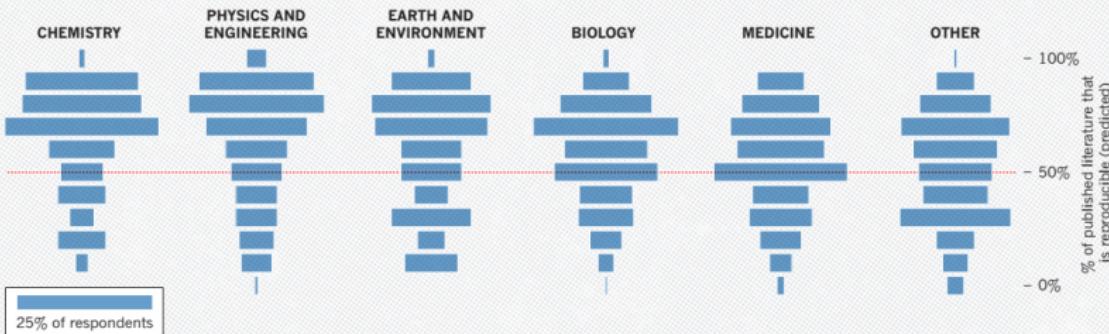
A Reproducibility Crisis (Baker 2015)

A 'CRISIS' IN NUMBERS

Nature surveyed 1,576 scientists online to get their thoughts on reproducibility in their field and in science in general. See go.nature.com/2vjr4y for more charts and access to the full data.

HOW MUCH PUBLISHED WORK IN YOUR FIELD IS REPRODUCIBLE?

Physicists and chemists were most confident in the literature.

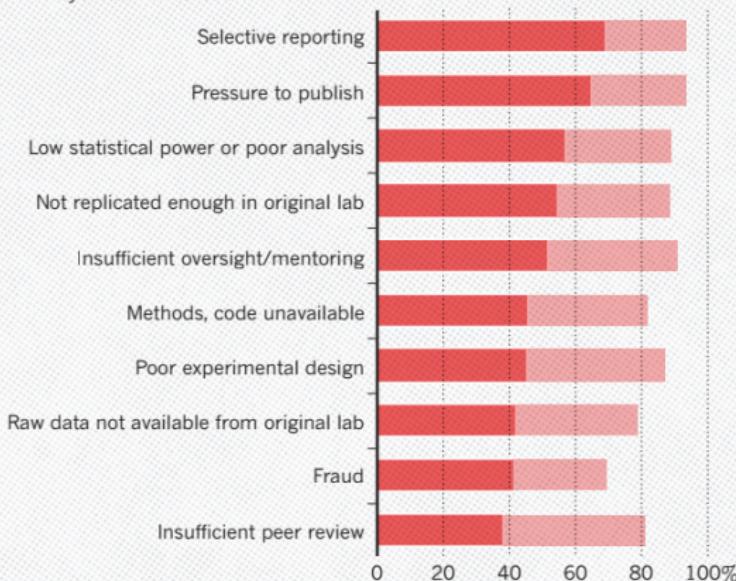


A Reproducibility Crisis (Baker 2015)

WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?

Many top-rated factors relate to intense competition and time pressure.

- Always/often contribute
- Sometimes contribute



Open Process (Pasquier et al. 2017)



SCIENTIFIC DATA

A graphic consisting of four rows of binary digits (0s and 1s) in blue. The first row is 1101110, the second is 01111101, the third is 110111110, and the fourth is 011101101.

OPEN

Comment: If these data could talk

Thomas Pasquier¹, Matthew K. Lau², Ana Trisovic^{3,4}, Emery Boose², Ben Couturier², Mercè Crosas⁵, Aaron M. Ellison², Valerie Gibson⁴, Chris Jones⁴ & Margo Seltzer¹

In the last few decades, data-driven methods have come to dominate many fields of scientific inquiry. Open data and open-source software have enabled the rapid implementation of novel methods to manage and analyze the growing flood of data. However, it has become apparent that many scientific fields exhibit distressingly low rates of repeatability and reproducibility. Although there are many dimensions to this issue, we believe that there is a lack of formalism used when describing end-to-end published results, from the data source to the analysis to the final published results. Even when authors do their best to make their research and data accessible, this lack of formalism reduces the clarity and efficiency of reporting, which contributes to issues of reproducibility. Data provenance aids both repeatability and reproducibility through systematic and formal records of the relationships among data sources, processes, datasets, publications and researchers.

Received: 12 April 2017

Accepted: 24 July 2017

Published: xx xxx 2017

Challenge: How do you know what the code actually does?

```
q <- runif(1)
if (q > 0.5){
    print("Yay!")
}else{
    print("booooo!")
}
```

Data Provenance and Ecological Networks

Methods in Ecology and Evolution

[Explore this journal >](#)

[View issue TOC](#)

Volume 5, Issue 11

November 2014

Pages 1206–1213

Application

enaR: An R package for Ecosystem Network Analysis

Stuart R. Borrett , Matthew K. Lau

First published:

27 October 2014 [Full publication history](#)

DOI:

10.1111/2041-210X.12282 [View/save citation](#)

Cited by (CrossRef):

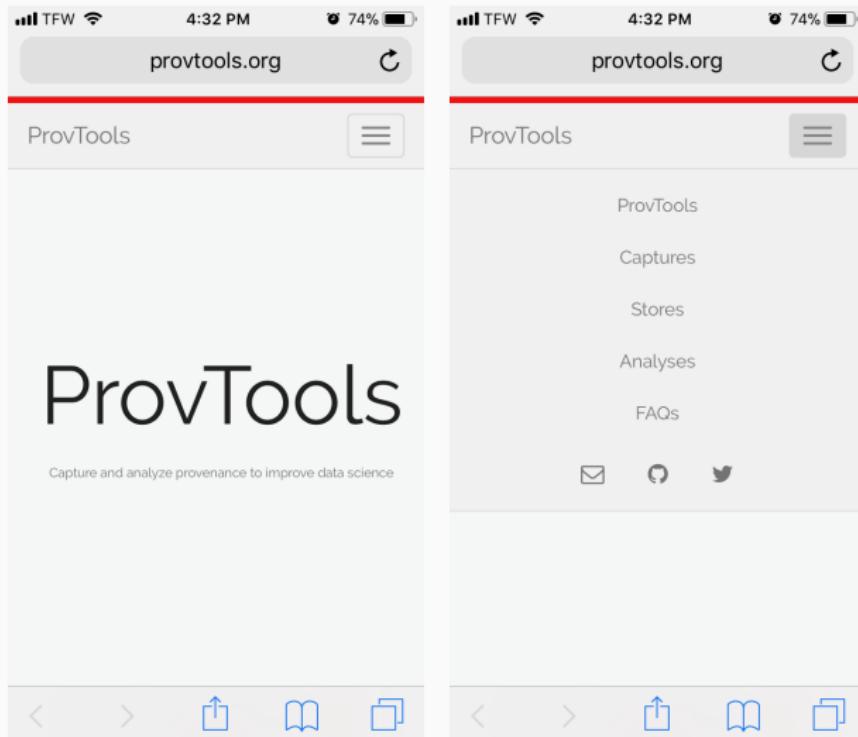
20 articles  |

 Citation tools ▾

ProvTools

Capture and analyze provenance to improve data science

ProvTools



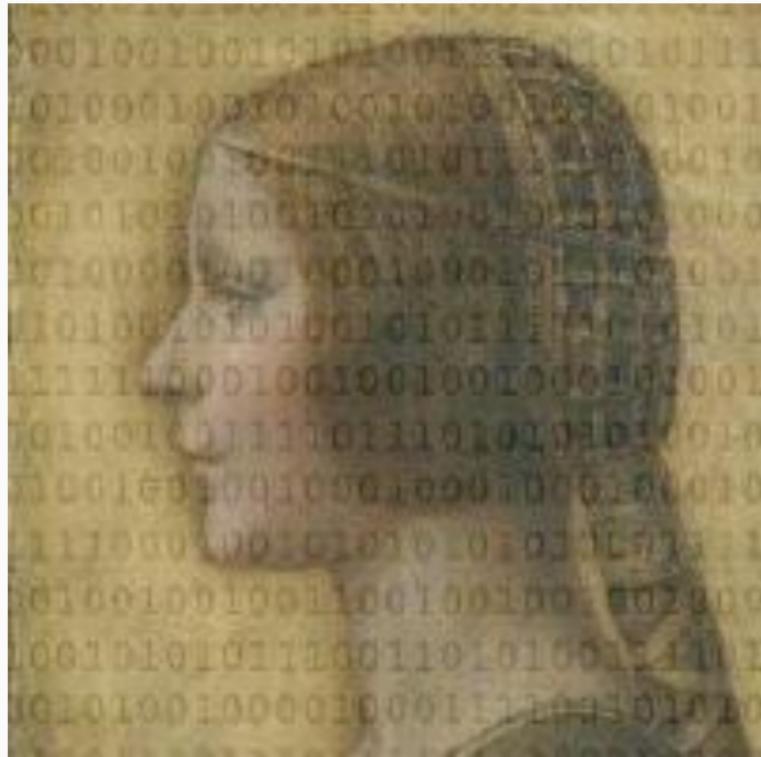
ProvTools

The image shows two side-by-side screenshots of a mobile web browser displaying the provtools.org website. Both screenshots have a top bar showing signal strength, time (4:32 PM), battery level (74%), and a refresh icon.

Screenshot 1 (Left): The main content area displays a sidebar menu with the following items: "ProvTools", "ProvTools", "Captures", "Stores", "Analyses", and "FAQs". Below the menu are social sharing icons for email, a circular icon, and Twitter. At the bottom are standard mobile navigation icons: back, forward, search, and a square.

Screenshot 2 (Right): The main content area has a red header bar with the URL "provtools.org". The page title is "Provenance". Below the title is a sub-section titled "What is it and why use it?". The main heading on the page is "What is provenance?". A descriptive paragraph follows: "Provenance is information about the processes that used and produced data." Below this is another heading, "Why use it?", with a detailed explanatory paragraph underneath.

provR (aka. RDataTracker)



cleanR

```
18 #### Some datasets are loaded and no
19 #### longer used.
20 #### Like this one
21 data.16.2 <- read.csv('projects/data_
22   forestplot/dataset_v2_june_from_
23   collaborator1.csv')
24
25 #### Create a bunch of intermediate
26 #### objects
27 data.v1.1to4 <- data.16[,1:4]
28 data.v1.1to4. <- data.v1.1to4
29 data.v1.1to4 <- data.v1.1to4 * 2
30 data.v1.1to4.2 <- data.v1.1to4 * 2
31 data.16[,1:4] <- data.v1.1to4.2
32 library('vegan')
33 d1 <- vegdist(data.16[,1:2])
34 d2 <- vegdist(data.16[,2:3])
35
36 #### Conduct some analyses
37 mant1 <- mantel(d1,d2)
38 mant2 <- mantel(d2,d1)
39 mant11 <- mantel(d1,d1)
40 f1t1 <- lm(Sepal.Length~Sepal.Width,data=
41   data.16)
42 lm.summary.1 <- summary(f1t1)
43
44 #### write some data to file
45 write.csv(data.v1.1to4,'projects/data_
46   _forestplot/savel.csv',row.names = F)
47
48 #### write some analyses to file
49 capture.output(lm.summary.1, file="
50   analysis_forest/anova_table_1.txt")
51
52 #### write some figures to file
53 #### Here's another random, unused package
54 library('txtplot')
55
56 png("figures_1/fig1_biplot.png")
57 plot(data.16[,1:2])
58 dev.off()
59
60 png("figures_1/fig1_biplot_t2.png")
61 plot(data.16[,1:2],2)
62 dev.off()
63
64 png("figures_2/fig2_biplot.png")
65 plot(data.16[,2:3])
66 dev.off()
```

Listing 7. Original “messy” code.

```
1 data.16 <- read.csv("projects/2016_
2   biomass_survey.csv")
3 data.v1.1to4 <- data.16[, 1:4]
4 data.v1.1to4 <- data.v1.1to4 * 2
5 data.v1.1to4.2 <- data.v1.1to4 * 2
6 data.16[, 1:4] <- data.v1.1to4.2
7
8 png("figures_2/fig2_biplot.png")
9 plot(data.16[, 2:3])
10 dev.off()
```

Listing 9. Curated code for figure 2

encapsulator

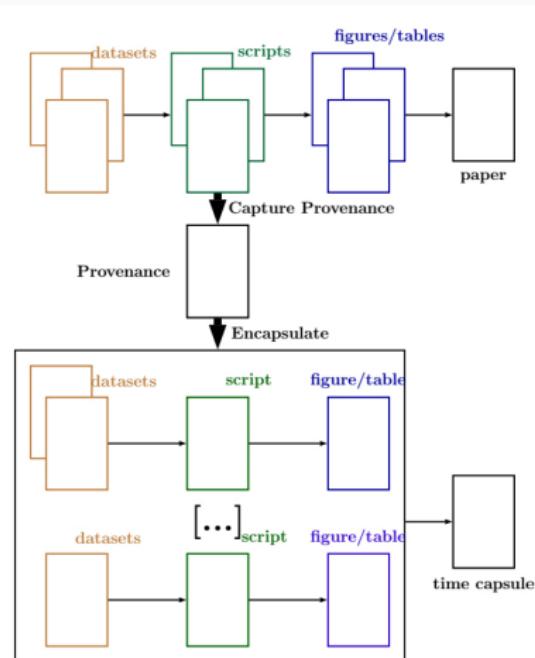


Figure 3. The encapsulation process.

enaR model packing (Borrett and Lau 2014)

1210 S. R. Borrett & M. K. Lau

```
library(enaR) # load enaR package
> # -- ENTER MODEL DATA -- from Dame and Patten (1981)
> # node names
> names <- c("Filter Feeders","Microbiota","Meiofauna",
+ "Deposit Feeders","Predators","Deposited Detritus")
> # Internal Flows of model, as matrix (oriented row to column)
> F <- matrix(c(0, 0, 0, 0, 0, 0, 0,
+ 0, 0, 8.1721, 0, 1.2060, 0, 0, 0, 7.2745,
+ 0, 1.2060, 0.6609, 0, 0, 0.6431, 0.5135, 0, 0,
+ 0.1721, 0, 0, 15.7910, 0, 4.2403, 1.9076, 0.3262, 0),
+ ncol=6)
> rownames(F) <- names # add node names to rows
> colnames(F) <- names # add node names to cols
> # boundary flows
> inputs <- c(41.47,0, 0, 0, 0, 0)
> outputs <- c(25.1650, 5.76, 3.5794, 0.4303, 0.3594, 6.1759)
> # Living
> Living <- c(TRUE,TRUE,TRUE,TRUE,TRUE,FALSE)
> # pack the model data into the R network data object
> m <- pack(flow=F,input=inputs, respiration=outputs, outputs=outputs, living=Living)
>
> ssCheck(m) # check to see if the model is at steady-state
[1] TRUE
> # perform flow analysis
> F <- enaFlow(m) # perform ENA flow analysis
> attributes(F) # show analysis objects created
$names
[1] "T"  "G"  "GP" "N"  "NP" "ns"
> F$ns # show flow analysis network statistics
  Boundary    TST TSTp   APL    FCI    BFI    DFI    IFI
[1,] 41.47 83.5833 NA 2.015512 0.1101686 0.4961517 0.1950689 0.3087794
     ID.F ID.F.I ID.F.O HMG.I HMG.O AMP.I AMP.O mode0.F mode1.F
[1,] 1.582925 1.716607 1.534181 2.051826 1.891638 3      1 41.47 32.90504
     mode2.F mode3.F mode4.F
[1,] 9.208256 32.90504 41.47
> F$T
   Filter Feeders       Microbiota       Meiofauna       Deposit Feeders
        41.4700          8.1721          8.4805          2.5100
   Predators Deposited Detritus
        0.6856          22.2651
```

Policy: OPEN Science = data + source + process

Ocean & Coastal Management 115 (2015) 71–76

 ELSEVIER

Contents lists available at [ScienceDirect](#)

Ocean & Coastal Management

journal homepage: www.elsevier.com/locate/ocecoaman



"Back off, man, I'm a scientist!" When marine conservation science meets policy



Naomi A. Rose ^a, E.C.M. Parsons ^{b,*}

^a Animal Welfare Institute, Washington DC, USA

^b Department of Environmental Science & Policy, George Mason University, Fairfax, VA, USA

ARTICLE INFO

Article history:

Received 4 January 2015

Received in revised form

24 April 2015

Accepted 27 April 2015

ABSTRACT

There is often a basic tension at the boundary between science and policy – the former seeks unbiased, objective descriptions of reality, while the latter must incorporate various factors in its development, including values, ideologies, economics, biases, and emotions. Problems may arise if, and when, marine scientists who enter the policy arena fail to understand these differing priorities, and we describe some

Policy: OPEN Science = data + source + process

Troublesome scientists

Here are three 'types' of scientists who can create difficulties (rather than assist in solutions) when involving themselves in policy debates and discussions.

The Naive Scientist

These scientists believe that if only policy-makers had the right information, they would make the right decisions. They do not understand the important human and legal dimensions of policy-making and make little attempt to interpret their work in that context. As one researcher stated in an interview, "*If all sides devoted their resources to research rather than to lawsuits, we could get some answers, but without them, the lawsuits will continue*" ([Madin, 2009](#)).

The 'Ivory Tower' Scientist

They believe that it is essential for scientists to remain 'pure', to stay at arm's length from anything resembling advocacy, even though they may also seek media attention for their work. This may result in essential data not reaching policy-makers, resulting in poor decisions. Or, even worse, their research is mischaracterized or misunderstood and they make no attempt to correct these misinterpretations (see below). As a researcher once told one of the authors (Rose): "*I can't be held responsible for the policy implications of my work.*"

The 'Industry' Scientist

They work directly for special interests, but expect their science (even when not peer-reviewed) to be accepted as objective. They either do not understand their conflicts of interest, or ignore them. As one government scientist emphatically stated in a policy meeting attended by one of the authors (Rose), in response to a comment from a participant that there was disagreement over industry research results, "*Only if you disagree with science!*"

Thank you

Interested in alpha testing or more information about provenance and ProvTools:

provtools.org

Email me: *matthewklau@fas.harvard.edu*